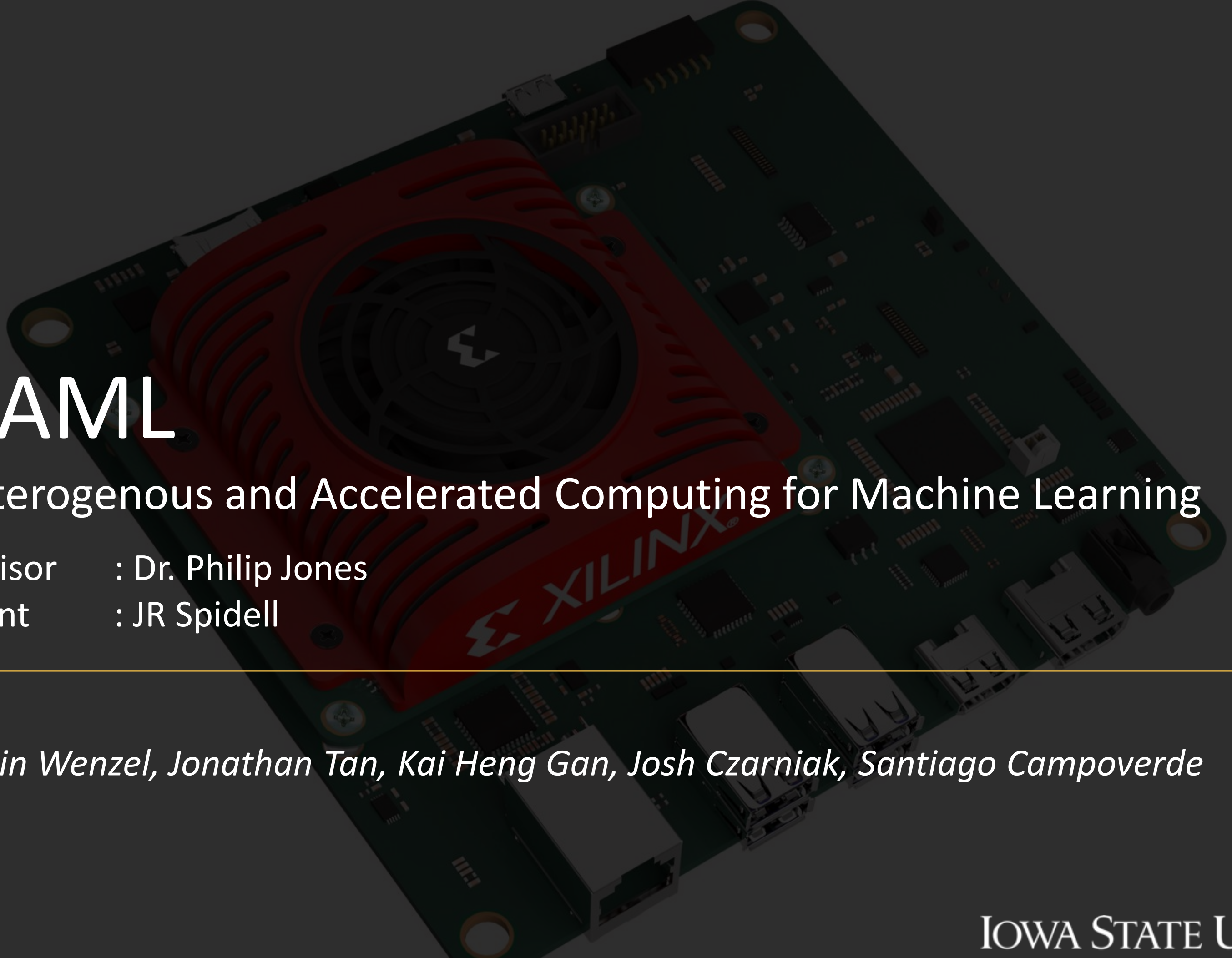


HAML

Heterogenous and Accelerated Computing for Machine Learning

Advisor : Dr. Philip Jones
Client : JR Spidell

Justin Wenzel, Jonathan Tan, Kai Heng Gan, Josh Czarniak, Santiago Campoverde

IOWA STATE UNIVERSITY

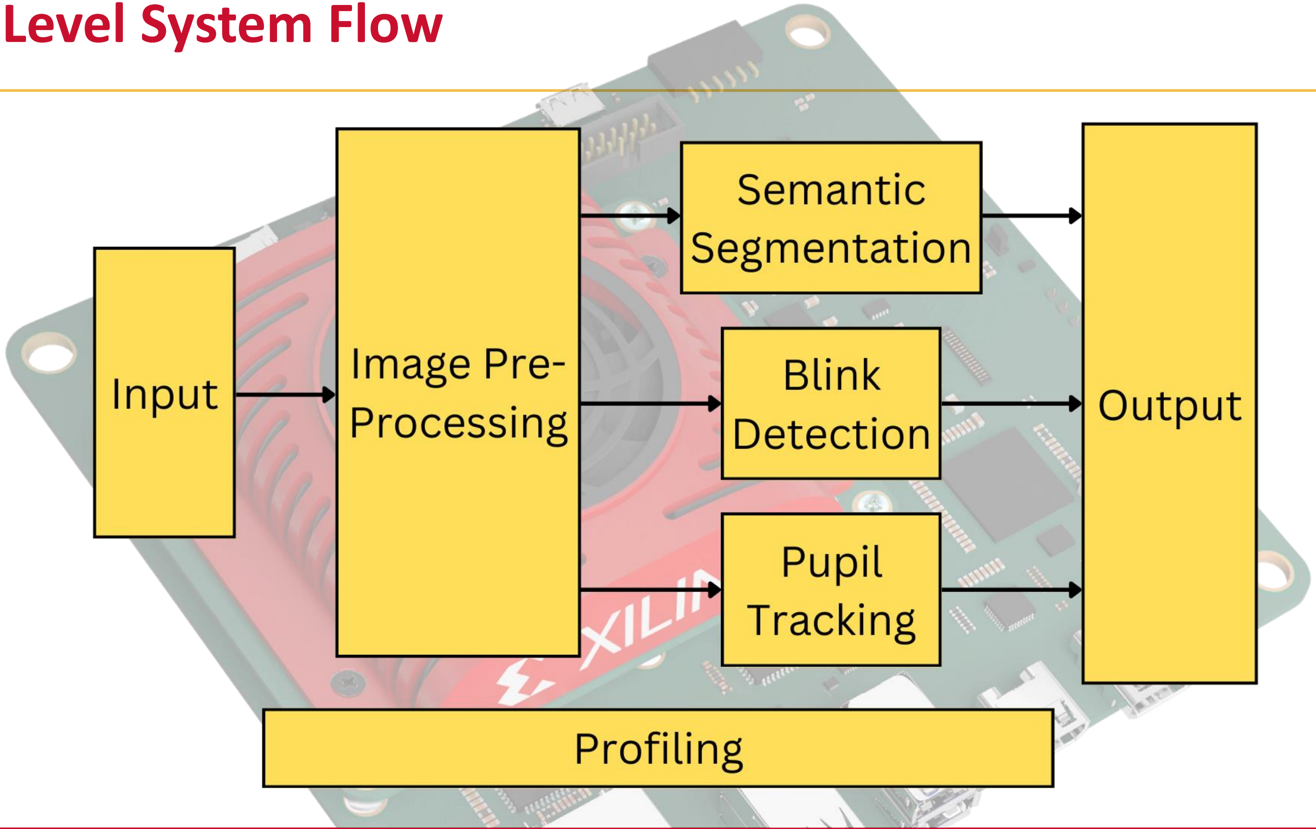
Problem Statement

- Client wants to create a wheelchair system to help people with disabilities complete day-to-day activities by tracking pupil movement.
  - Using pupil movement to control mouse cursor
  - Prediction of the user’s state (predicting seizures, stress, fatigue, etc.)
- Three different models crucial for the success of this goal:
  - Blink Detection
  - Pupil Tracking
  - Semantic Segmentation
- Our project is:
  - A subcomponent of a larger ML powered wheelchair system.
  - Part of a series of other senior design groups.

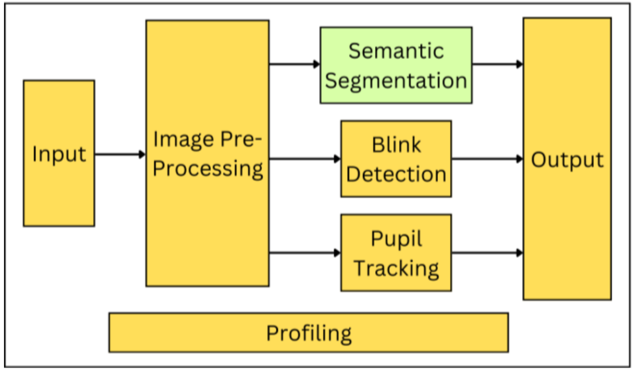
Functional Requirement and Constrains

- Functional Requirement:
  1. Create a system with **3 models** (blink detection, pupil tracking, semantic segmentation) running in parallel.
     - **No timing requirements**.
     - Achieve semantic segmentation accuracy of 90%
  2. Create a system with blink detection and pupil tracking running in parallel and achieve **throughput of 200 FPS**.
- Constraints:
  - Client provides 2 ML models (blink and pupil tracking)
  - Client wants it implemented on the Xilinx Kria KV260 evaluation board

High Level System Flow



Semantic Segmentation



Purpose:

- Remove glare from iris images

Model Type:

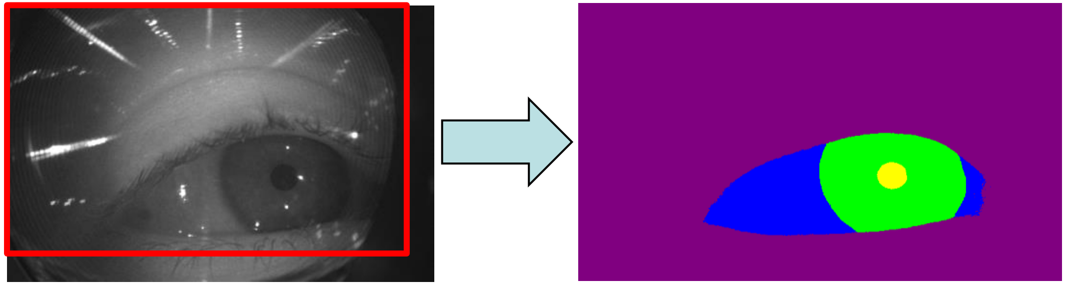
- Pixel-wise classification model

Input:

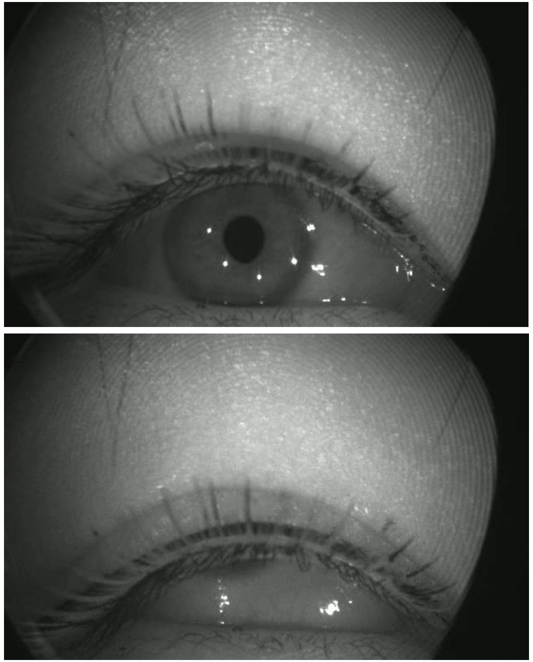
- Frames extracted from a given video
- 1-channel image (grayscale)

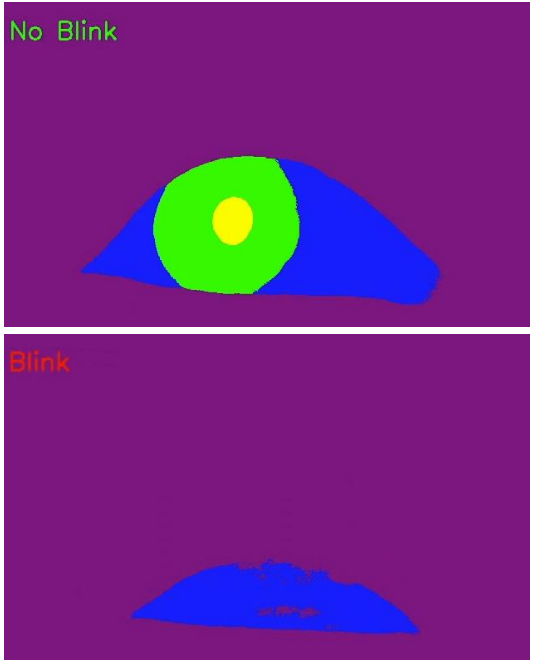
Output:

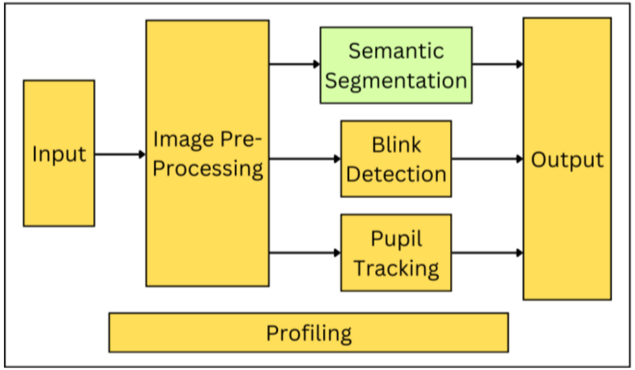
- 4-channel segmented image
- Class indices array



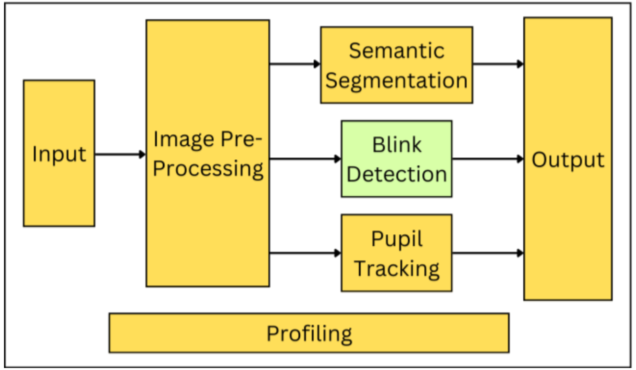
Semantic Segmentation







Blink Detection



Purpose:

- Detect blink in a frame.

Input:

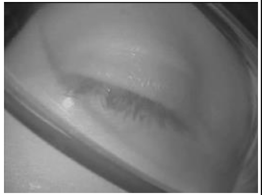
- Frames extracted from a given video

Model Type:

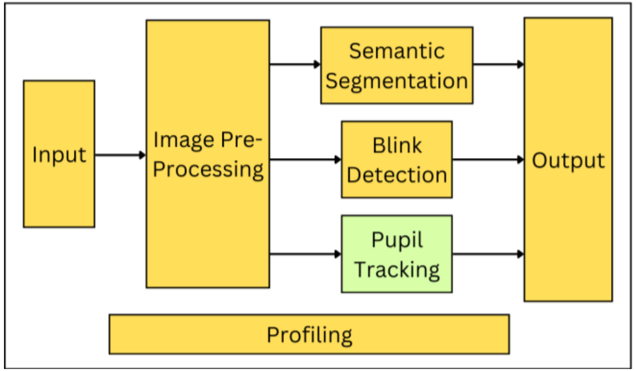
- Classification model

Output:

- Two classes of classification:
  - ❖ Blink: "Frame contains a blink"
  - ❖ No Blink: "Frame does not contain a blink"
- Neural Network outputs the probability that a video frame is a blink or no blink

	Input Image
No blink	
Blink	

Pupil Tracking



Purpose:

- Track pupil coordinates in a frame.

Input:

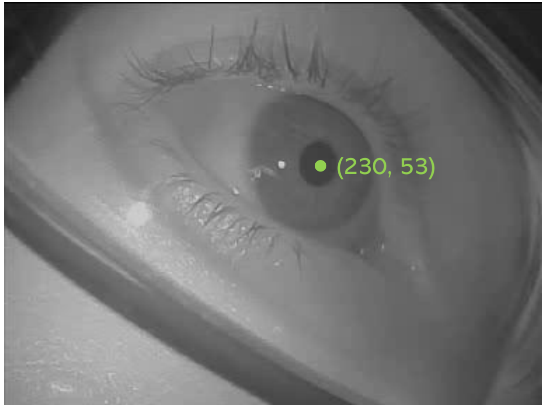
- Frames extracted from a given video

Model Type :

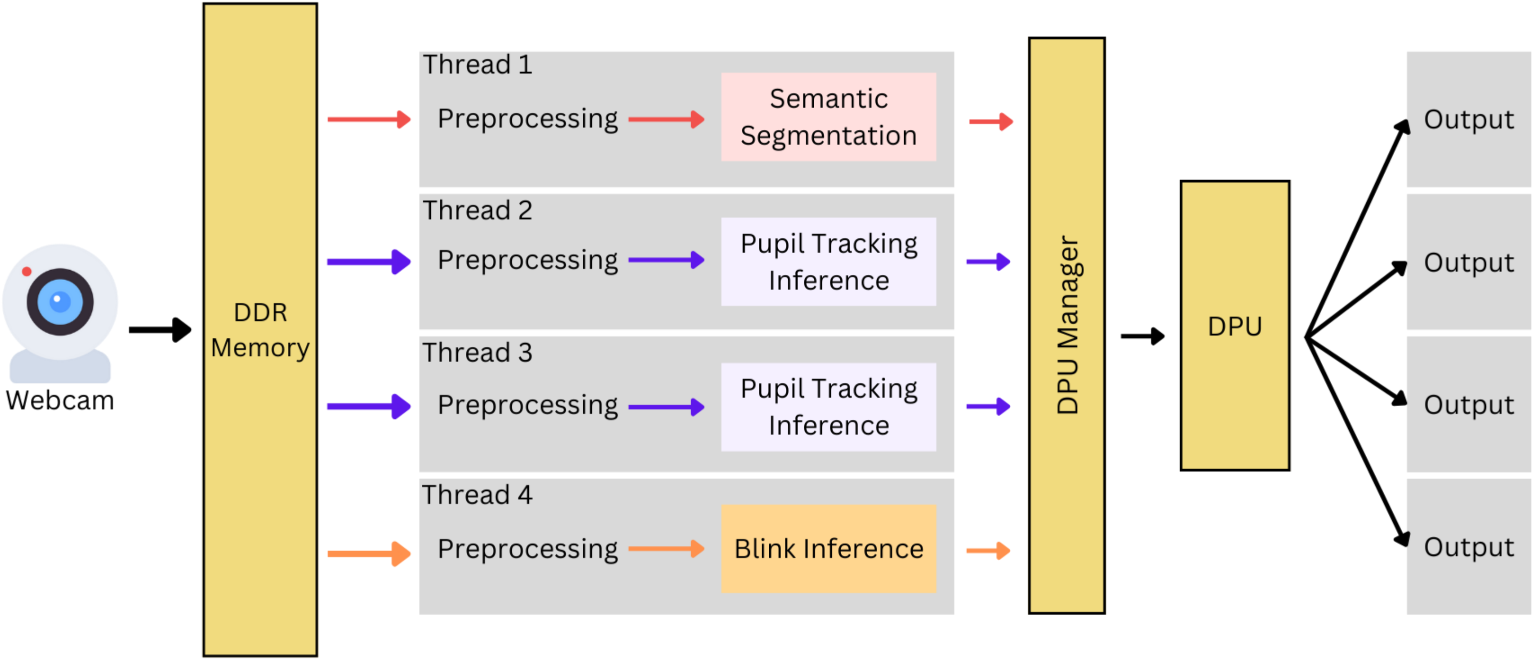
- Regression model

Outputs:

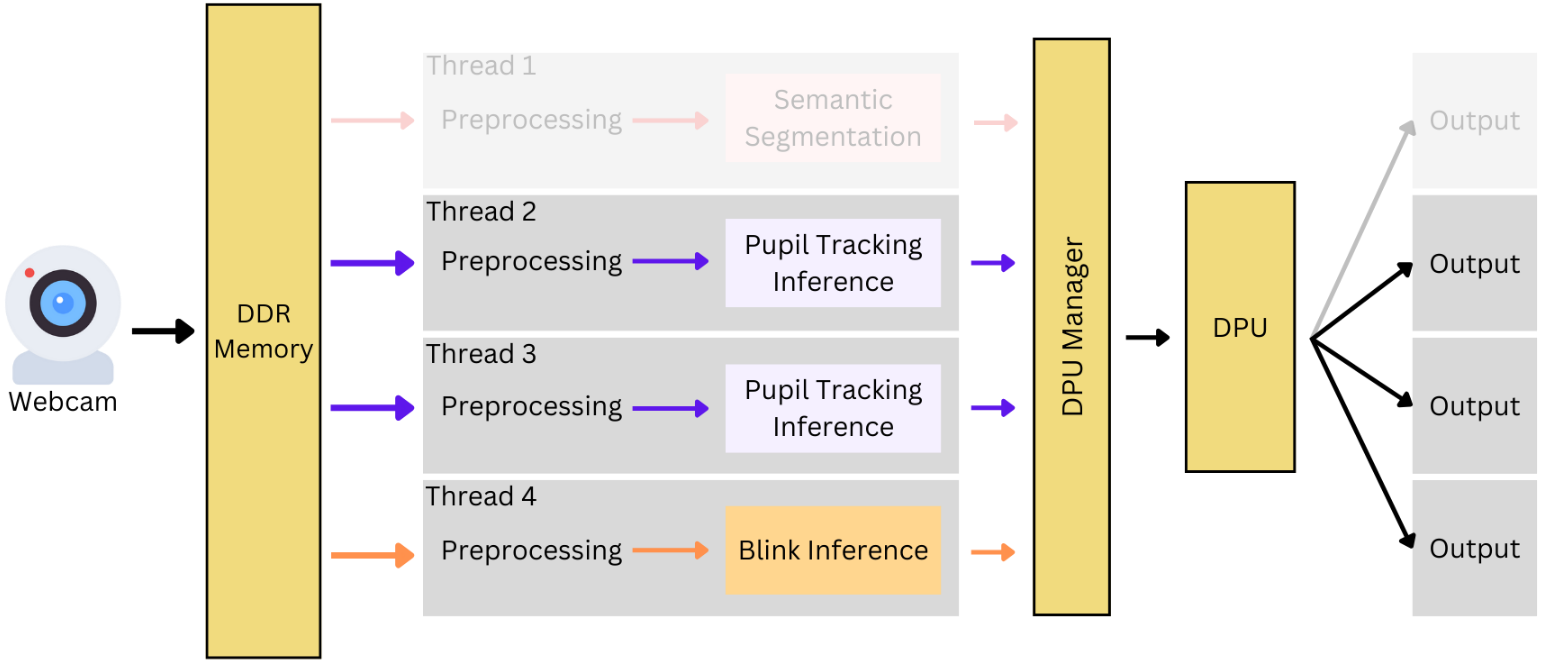
- Returns the location of the pupil within the image frame
- Given in X and Y coordinates using pixel measurements
- Slower run time than blink model



Multithreaded Application – Requirement 1



Multithreaded Application – Requirement 2



DPU Resource Management

DPU

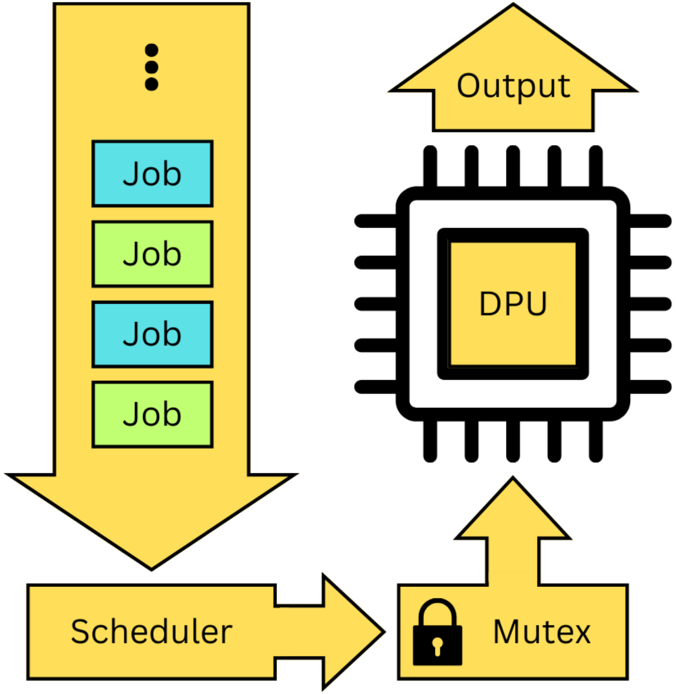
- Special processor for performing CNN operations.
- Only one fits on the Kria FPGA.

DPU Manager

- Schedule access to the DPU

How does sharing work?

- Scheduler: First-come-first-serve method.
- Mutex locks: Prevent access to DPU when it is busy, thereby only allowing one thread to use the DPU at one time.



Results – Requirement 1 (tri-model integration)

- Requirement 1:
  - Create a system with 3 models (blink, pupil tracking, semantic segmentation) running in parallel.
  - No timing requirements.
  - Create the semantic segmentation model.
- Results (throughput):
  - Single threaded program : 10.25 FPS
  - Multithreaded program : 10.83 FPS
- Comments:
  - Semantic segmentation has significantly higher latency.

Results – Requirement 2 (throughput)

- Requirement 2:
  - Create a system with blink detection and pupil tracking running in parallel.
  - Achieve throughput of 200 FPS.
- Results (throughput):
  - Single threaded program : 16 FPS
  - Multithreaded program : ≈ **200 FPS**

Model Accuracy Analysis

Semantic Segmentation Testing

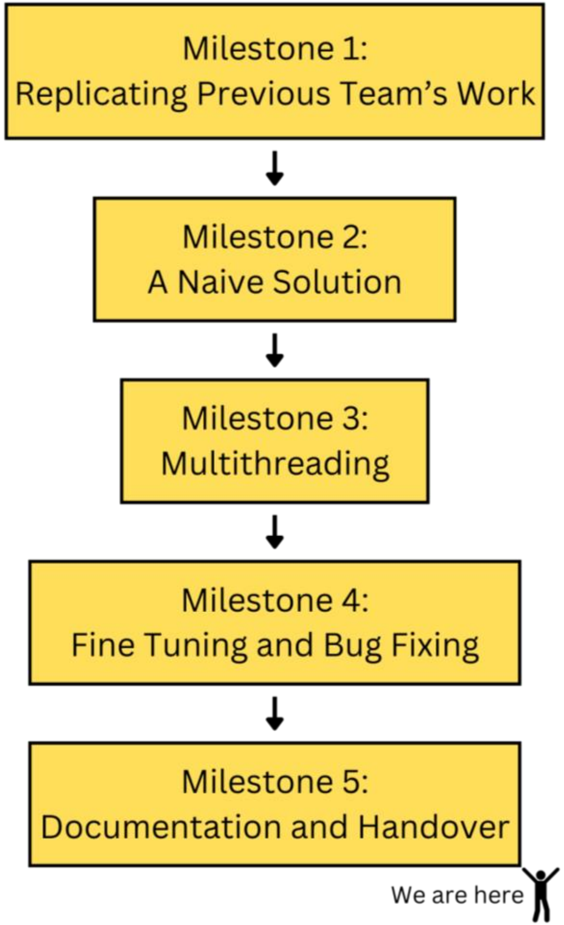
- Method: Mean Intersection Over Union
- Used to effectively distinguish a class area on an image and allows for a pixel precision comparison
- Accuracy: $\approx$ 98%

Blink Detection Testing

- Method: Confusion Matrix
- Allows for a full coverage analysis between two binary values and provides extra data on True Positives/Negatives and False Positives/Negatives

Pupil Tracking Testing

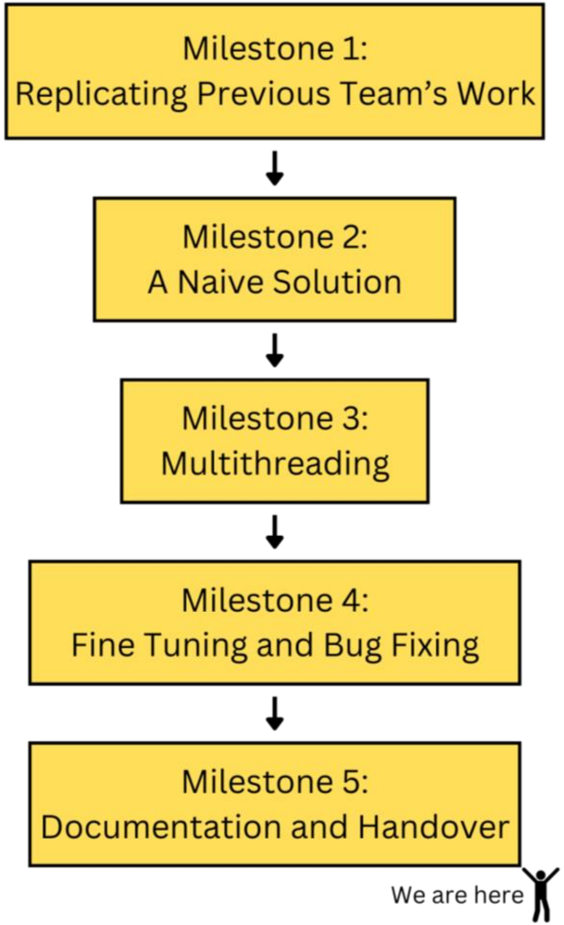
- Method: Root Mean Squared Error
- Helps identifying the boundary of average errors on the prediction set, allowing for a clear performance indicator



Project Milestone & Schedules

- Milestone 2:
  - Implementing a **single threaded** approach
  - Identify bottlenecks and overheads through profiling
- Milestone 3:
  - Implement a **multithreaded** approach
  - Measure model accuracy
  - Perform timing analysis
- Milestone 4:
  - Fine tune milestone 3 code, fix bugs
  - Refining semantic segmentation model

Conclusion



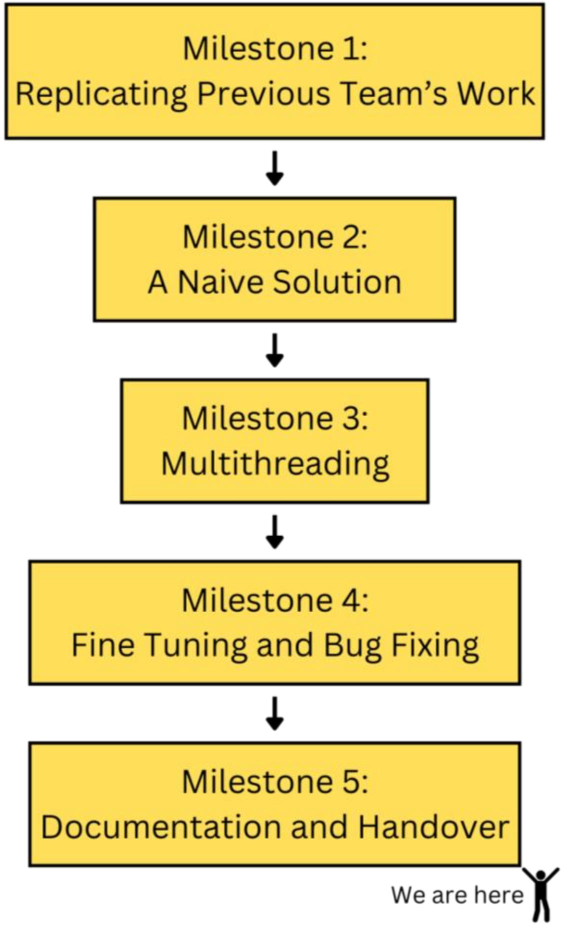
Senior design achievements:

- Successfully implement all models (semantic segmentation, blink, pupil tracking) onto the Xilinx Kria KV260 board.
- Successfully implement multithreaded program.
- Successfully met throughput of 200 FPS.
- **Met** client and advisor requirements and expectations.

Questions?

Senior design achievements:

- Successfully implement all models (semantic segmentation, blink, pupil tracking) onto the Xilinx Kria KV260 board.
- Successfully implement multithreaded program
- Successfully met throughput of 200 FPS
- **Met** client and advisor requirements and expectations



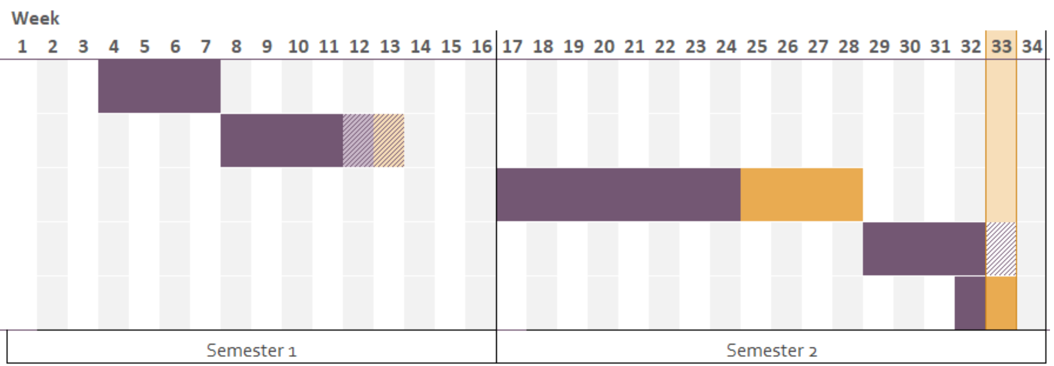
Questions

1. Grant chart
2. DPU
3. BRAM
4. Cyber security implications
5. Broader context
6. Potential improvements
7. Mean intersection over union
8. RMSE
9. Confusion matrix

Grant Chart

Current period: 33 | Plan Duration | Actual Start | % Complete | Actual (beyond plan) | % Complete

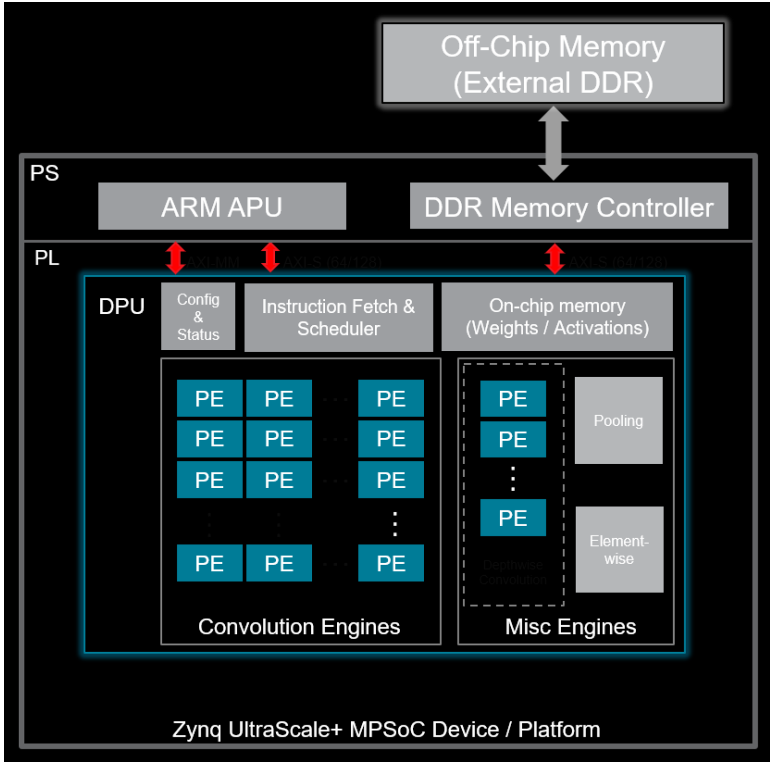
ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Milestone 01	4	4	4	4	100%
Milestone 02	8	5	8	6	80%
Milestone 03	17	8	17	12	100%
Milestone 04	29	5	29	4	100%
Milestone 05	32	1	32	2	100%



What is the DPU?

Deep-learning **P**rocessing **U**nit

- A programmable engine for convolutional neural network
- An IP block in Vivado, ours is the B4096 architecture

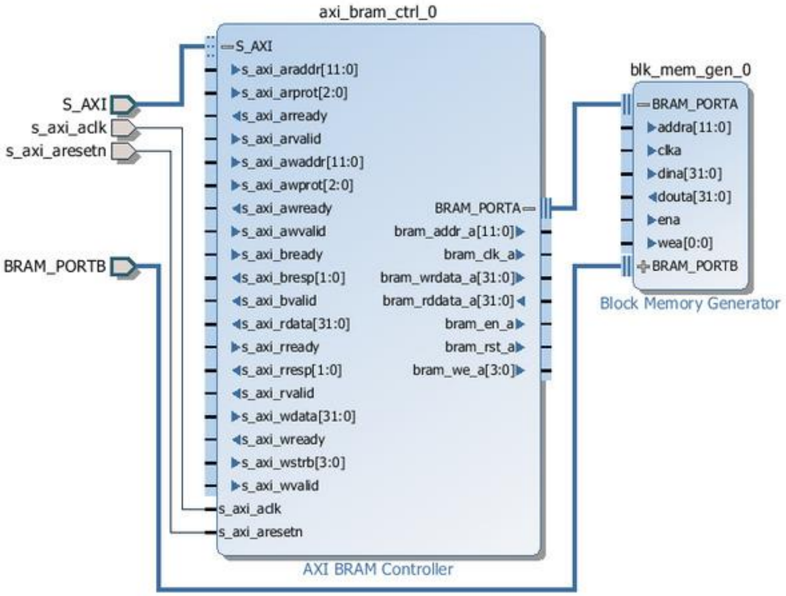


“DPU For Convolutional Neural Network.” AMD, www.xilinx.com/products/intellectual-property/dpu.html.

Return to Questions

What is BRAM and why is it important?

- **B**lock **R**andom **A**ccess **M**emory
- A type of memory used in FPGAs
- High-speed data access
- In our project, it is used by the DPU



AMD. support.xilinx.com/s/question/0D52E00006hpZsESAU/axi-bram-controller-unable-to-change-address-to-least-significant-bits?language=en_US.

Return to Questions

Cyber security implications?

Although the system will mostly be used "off the grid," some security implications are worth paying attention to:

- Intrusion during firmware update
- Install malware or ransomware on the board accidentally
- Physical security risks: Attacker installing malware physically

Return to Questions

Broader Context

Area	Description
Public health, safety, and welfare	• Bring an accessibility solution to people with disabilities.
Environmental	• ML application can be energy intensive. • While meeting functional target is important, we don’t rule out optimizing resources used by our system.
Global	• Opportunities for a machine vision application is endless, allowing huge improvements to human’s day-to-day life.
Economic	• The economic factors involve the cost of deployment, development, and the potential market demand which will influence future production costs and scale.

Return to Questions

DNU! What are some potential improvements? (Outdated)

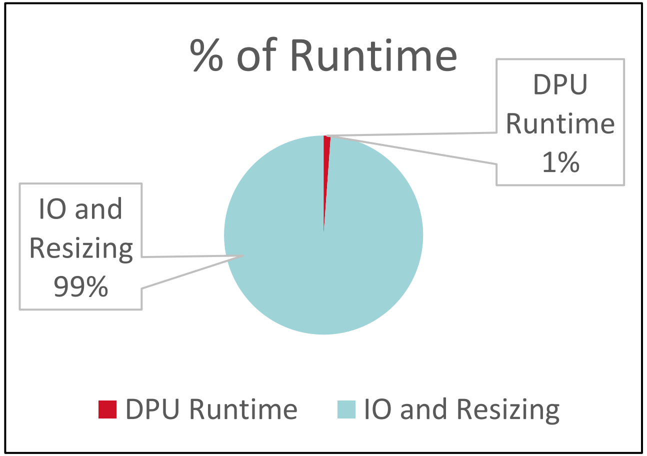
Throughput of Naïve solution: 16FPS

- Single threaded
- Redundant file read (reading same input twice)
- Inefficient image resizing technique
- Using Amdahl's Law, if we remove redundant file read:

$$Speedup = \frac{T_{org}}{\left((1-f)+\frac{f}{a}\right) \times T_{org}} = \frac{1}{(1-f)+\frac{f}{a}}$$

$$= \frac{1}{(1-0.99)+\frac{0.99}{2}} = 1.98$$

- Assumption: Both Blink and Pupil Tracking algorithm perform same IO read and resizing.



What are some potential improvements?

- Multicore DPU
- Optimized semantic segmentation

Return to Questions

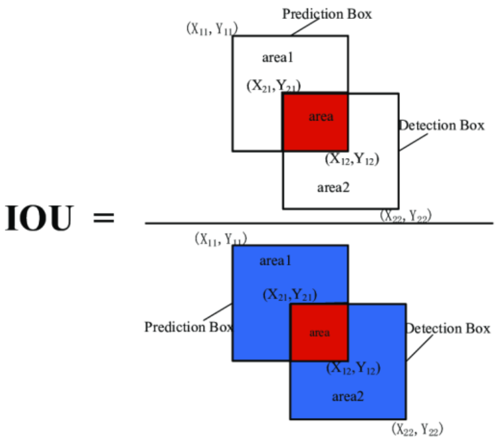
What is mean intersection over union (IoU)?

Mean Intersection over Union (IoU):

- Average IoU across all classes in multi-class segmentation tasks.
- Reflects overall segmentation accuracy, normalized between 0 (no overlap) and 1 (perfect overlap).

Advantages of mIoU:

- Normalizes performance across imbalanced class sizes.
- Widely used metric for comparing model effectiveness in benchmarks.



[Return to Questions]

RMSE

RMSE calculates the square root of the average squared differences between the predicted and actual values.

- Square to make sure all numbers are positive, and errors are bigger
- Add all values up and then divide by the number of predictions
- Square root the number to bring it back into the original measurement range

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

- n = number of observations
- y_i = observed values
- $\hat{y}_i$ = predicted values

Return to Questions

Confusion Matrix

- A binary classification Confusion Matrix utilizes a 2x2 table
  - **True Positive (TP)**: The number of cases correctly predicted as positive.
  - **True Negative (TN)**: The number of cases correctly predicted as negative.
  - **False Positive (FP)**: The number of cases incorrectly predicted as positive (also known as Type I error).
  - **False Negative (FN)**: The number of cases incorrectly predicted as negative (also known as Type II error).

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- Accuracy equation: $\frac{(TP+TN)}{(TP+FP+FN+TN)}$

Return to Questions